Fanpu Cao

fanpucao@gmail.com · (+86) 189-9894-1792 · Github

Education

South China University of Technology (SCUT), Guangzhou, China Sep. 2021 – Present

Undergraduate student majoring in Artificial Intelligence (AI)
GPA: 3.72/4.0 *(Top 25%)*
Languages: *English* - IELTS **7.0**, CET6 Grade **550**, CET4 Grade **602**.
Research Interests: vLLM & LLMs, Model acceleration, Long-term time series forecasting, AI in FinTech, Embodied AI.

Core Courses

- C++ Programming (4.0/4.0)
- Python Programming (4.0/4.0)
- Advanced Language Programming Practice (4.0/4.0)
- Machine Learning (4.0/4.0)
- Deep Learning and Computer Vision (4.0/4.0)
- Big Data and Data Mining (4.0/4.0)
- Optimization Methods (4.0/4.0)
- Large Language Models and AI Engineering Design (4.0/4.0)

Papers

Ister: Inverted Seasonal-Trend Decomposition Transformer for Explainable Multivariate Time Series Forecasting March. 2024 – July. 2024

Advised by Dr. Shu Yang, Next Generation Internet R&D Center, Tsinghua University Research Institute, Shenzhen. Research on tasks associated with time series analysis, long-term forecasting. In long-term time series forecasting, Transformer-based models have achieved great success, due to its ability to capture long-range dependencies. However, existing Transformer-based models struggle to accurately identify the critical components of time series for prediction, limiting both model interpretability and predictive accuracy. Besides, it faces scalability issues due to quadratic computational complexity of self-attention. In this paper, we propose a new model named *Inverted Seasonal-Trend Decomposition Transformer (Ister)*, which addresses these challenges in long-term multivariate time series forecasting by designing an improved Transformer-based structure. Under review by *IJCAI 2025*. Arxiv: https://arxiv.org/abs/2412.18798.

CEDTS-RL: Towards efficient and green cross-geographical data centers based on reinforcement learning 2024

Working with Dr. Shu Yang, Next Generation Internet R&D Center, Tsinghua University Research Institute, Shenzhen. Aiming at the huge energy consumption of data centers and the carbon emissions they generate, we propose a novel cross-data center task scheduling mechanism that aims to reduce the overall carbon emissions while satisfying user demands. We first constructed a task scheduling model designed to optimize carbon emissions while considering latency and energy consumption. Subsequently, a reinforcement learning-based algorithm called CEDTS-RL (Carbon Emission-Driven Task Scheduling Algorithm Based on Reinforcement Learning) is developed, which utilizes information on the production and consumption of renewable energy between different data centers.The effectiveness of the CEDTS-RL algorithm is validated through a comprehensive simulation using data center location data from commercial companies and NASA's surface climate dataset. The simulation results show that the proposed algorithm effectively reduces carbon emissions compared to the baseline algorithm, albeit with a slight increase in average latency. This work contributes to optimizing the environmental sustainability of data center operations and optimizing task allocation based on renewable energy availability. *Under review by IWQoS 2025*.

SWIFT: Mapping Sub-series with Wavelet Decomposition Improves Time Series Forecasting
March. 2024 – present

Co-authored papers. In recent work on time-series prediction, Transformers and even large language models have garnered significant attention due to their strong capabilities in sequence modeling. However, in practical deployments, time-series prediction often requires operation in resource-constrained environments, such as edge devices, which are unable to handle the computational overhead of large models. To address such scenarios, some lightweight models have been proposed, but they exhibit poor performance on non-stationary sequences. In this paper, we propose SWIFT, a lightweight model that is not only powerful, but also efficient in deployment and inference for Long-term Time Series Forecasting (LTSF). Under review by *IJCAI 2025*. Arxiv: https://arxiv.org/abs/2501.16178.

INTERNSHIPS

China Merchants Lion Rock Artificial Intelligence Laboratory Sep. 2024 – present

Work in the direction of embodied intelligence under the leadership of Dr. Jiaxing Zhang, Chief Scientist. The work consists of performance testing, paper reading and reproduction of papers on a variety of existing SOTA multi-modal models (eg. Open-Sora, Transfusion, Emu3, RT-H, etc.).
Our goal is:

- train a multi-modal vLLM that supports video and 3D image understanding, target detection, and tracking.
- implement generalization and scaling law for Robotics Reasoning based on the idea of trajectory prediction for operating robots + LLM.

Next Generation Internet R&D Center, Tsinghua University Research Institute, Shenzhen.
March. 2024 – Jun. 2024

Under the leadership of Dr. Shu Yang, participated in the development of China National Offshore Oil Corporation (CNOOC) Energy Large Model Project. Predicting energy data through Transformer-based models.

♡ HONORS AND AWARDS

Excellence Awards, Baidu Paddle Paddle Cup Dec. 2021
Outstanding Students, South China University of Technology-Baidu Pinecone Elite Class Jun. 2023
S Awards, 2024 Mathematical Contest In Modeling April. 2024
3rd Awards, 2023 Asia and Pacific Mathematical Contest in Modeling Nov. 2023

SKILLS

- Programming Languages: Python*, C++
- Maths: Calculus, Linear Algebra, Complex Functions, Probability&Statistics, Convex Optimization, Matrix analysis and calculation
- Frameworks: Pytorch , Pandas , GBDTs (lightgbm etc.), Visualization (Matplotlib, Seaborn, shap, etc.), Scikit-learn, Numpy, Scipy, OpenCV
- Platform&Tools: Windows, Linux, LaTeX, Markdown, Git, VSCode.
- Fields: CV(OpenCV, VGG, Resnet, YOLOv8, ViT), NLP(Transformers, Berts, LLMs), GANs, GNN, Data-Mining&Machine learning(Linear regression, SVM, KMeans, GBDTs, etc.)

PROJECT EXPERIENCE

Five-in-a-Row Game Development May 2022

Python, tensorflow Individual Project

Project link: https://github.com/macovaseas/five-in-a-row-

- Developed a five-in-a-row game using Python
- Supported local human vs. AI and human vs. human game-play
- Designed a simple and elegant game interface
- Implemented AI game-play using a CNN model

Texas Hold'em Game Development Dec 2022 – Jan 2023

Python Team Project

Project link: https://github.com/frinkleko/RL-Poker

- Developed a 1v1 Texas Hold'em game using OOP programming principles
- Designed a simple AI opponent based on mathematical principles
- Future improvements include designing multiplayer game-play and enhancing AI with reinforcement learning

SCUT-Baidu Pinecone Elite Program July 13, 2022 – May 31, 2023

Paddle Paddle aistudio Competition

- NLP Chinese news topic classification (**3rd/27**) using *Bert* pre-trained model
- CV classification competition (**5th/33**) using *resnet152* pre-trained model + random cropping + pseudo-labeling
- Awarded "Baidu Pinecone Talent Development Elite Program Outstanding Student"

Stock Price Prediction Nov 2022 – Dec 2022

torch, Deep learning Deep Learning Project

Using GRU network model, predicted stock prices for the Shanghai and Shenzhen 300 index from Jan 4, 2005, to Dec 31, 2019 (including open price, high price, low price, close price, price change percentage, trading volume, and trading value), and forecasted the stock prices from 2020 to Oct 28, 2022.

- Cleaned and preprocessed stock price data, predicted future stock prices based on historical data
- Built and trained a GRU model using the torch framework
- Formulated investment strategies based on the stock price predictions

Large Language Model Deployment and Fine-Tuning Oct 2023 – Nov 2023

LLM AIGC Project

Deployed ChatGLM-6B model on a server, fine-tuned the model on a custom dataset using the P-tuning-v2 algorithm, and achieved significant improvements in rouge and bleu scores compared to baseline.

- Deployed the model on a Linux terminal
- Fine-tuned the model using training scripts with custom hyperparameters
- Validated the fine-tuning results using validation scripts

Life Expectancy Nov 2023 – Dec 2023

Scikit-learn, Lightgbm Data Mining Project

Project link: https://github.com/frinkleko/Life-expectancy

- Used data mining techniques to extract key and interesting information from the Life Expectancy Data.csv
- Cleaned and preprocessed the data, analyzed data features (such as correlation, KDE distribution, data visualization results)
- Predicted life expectancy using LightGBM
- Visualized model interpretability using the shap library to identify features most related to life expectancy
- Conducted feature engineering using the RFECV algorithm to further enhance model prediction performance

Tree Drawing Suicide Project Oct 2023 – Dec 2023

Graph Neural Network, CNN Deep Learning Project

Guided by Prof. Ye Liu, this project uses a tree drawing dataset from Beijing Normal University, which includes tree drawings by children affected by disasters and associated suicide tendency labels. The project aims to build a deep learning model to classify and predict potential suicide tendencies based on children's drawings.

- Performed random oversampling to balance the classes
- Constructed a knowledge graph based on the characteristics of the tree drawings and psychological features
- Integrated CNN full convolution architecture with graph neural networks using Knowledge-Embedded Representation Learning, combined with gating mechanisms for prediction
- Split the data into test, validation, and training sets, achieving an accuracy of 97% on the validation set

- Used GradCAM for model heatmap visualization to identify features and elements in the drawings influencing suicide tendencies

IDEA-King May. 2024 – Jun. 2024

LLM Collaborative Projects

- Our GitHub project is available at this link. This project is the first to explore the capabilities of large language models (LLMs) in scoring research papers and generating ideas.
- The LLama-factory framework offers tools and examples for utilizing LLMs effectively. Users can configure the framework and base model following the provided documentation. This project uses Qwen1.5-7b-chat as the base model and example.
- We use the Alpaca format of the openreview dataset, which we have open-sourced. Users should configure the dataset according to the requirements of the LLama-factory project.
- To train the king model, we use the Proximal Policy Optimization (PPO) algorithm.